CENTER FOR PURE AND APPLIED MATHEMATICS
UNIVERSITY OF CALIFORNIA, BERKELEY

PAM- 635

A WAY TO FIND THE MOST REDUNDANT EQUATION IN A TRIDIAGONAL SYSTEM

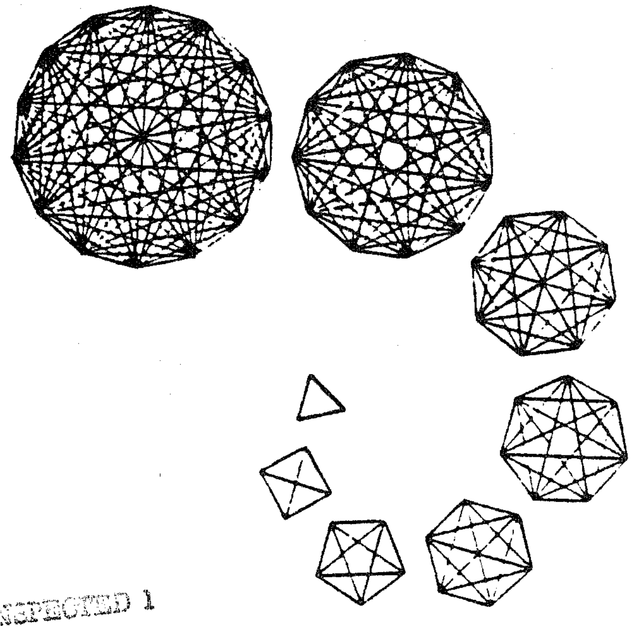
K. Vince Fernando

Beresford N. Parlett

and

Inderjit S. Dhillon

May 1995



This report was done with support from the Center for Pure and Applied Mathematics. Any conclusions or opinions expressed in this report represent solely those of the author(s) and not necessarily those of the Center for Pure and Applied Mathematics or the Department of Mathematics.

A Way to Find the Most Redundant Equation in a Tridiagonal System

K. Vince Fernando [*] Beresford N. Parlett [†]
and Inderjit S. Dhillon [‡]

April 26, 1995

Abstract

Suppose that one knows a very accurate approximation σ to an eigenvalue λ of a symmetric tridiagonal matrix T. A good way to approximate the eigenvector $\boldsymbol{x}$ is to discard an appropriate equation, say the rth, from the system $(T-\sigma I)\boldsymbol{x} = \mathbf{0}$ and then to solve the resulting underdetermined system in any of several stable ways. However the output $\boldsymbol{x}$ can be completely inaccurate if r is chosen poorly and in the absence of a quick and reliable way to choose r this method has lain neglected for over 35 years.

We show how double triangular factorization (down and up), which is closely related to 'twisted factorization', gives us directly the redundancy of *each* equation and so reveals the set of good choices for r.

The results extend to band matrices and the applications go beyond eigenvector computation to determinant evaluation and solution of well conditioned systems.

[*] NAG Ltd, Wilkinson House, Jordan Hill, Oxford OX2 8DR, UK.

[†] Mathematics Department, University of California, Berkeley, CA 94720, USA. Supported by ONR, contract N000014-90-J-1372.

[‡] Computer Science Division, EECS Department, University of California, Berkeley, CA 94720, USA. The author was supported in part by NSF Grant ASC-9313958, DOE Grants DE-FG03-94ER25219 and DE-FG03-94ER25206, Arpa contact DAAL03-91-C0047 via a subcontract from the the University of Tennessee, and by NSF under Infrastructure Grants CDA-8722788 and CDA-9401156.

1 Introduction

The task that started these investigations is the computation of eigenvectors of a symmetric tridiagonal matrix (entry (i,j) vanishes if $|i-j|>1$) once the eigenvalues are in hand. This is not a new problem and there are good programs available in libraries such as LAPACK and NAG. Nevertheless the experts do not consider the situation satisfactory, see [9]; the complexity of the programs seems out of proportion to the difficulty of the task and the adaptation of the current versions of inverse iteration to parallel mode is frustrating.

Let us briefly sketch the situation. Given an accurate approximation σ to an eigenvalue λ of an $n \times n$ symmetric tridiagonal matrix T one considers the solution $\boldsymbol{x}$ to the system of equations

$$(T-\sigma I)\boldsymbol{x} = \boldsymbol{b} \tag{1}$$

where $\boldsymbol{b}$ is to be chosen wisely. Since $\sigma \neq \lambda$ the best choice for $\boldsymbol{b}$ is the eigenvector we seek but this is not an option. Next best is to choose for $\boldsymbol{b}$ a column r of the identity matrix $I=(\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n)$. As will become clear in Section 3 choosing $\boldsymbol{b}=\boldsymbol{e}_r$ is equivalent to omitting Equation r from the system (1). The value $r=n$ was proposed by Wallace Givens in 1954, see [5], but no fixed value of r, independent of σ and T, will do.

Here is a quotation from Wilkinson concerning the computation of an eigenvector $\boldsymbol{u}_k$, in Chap. 5, Section 50, below Equation (50.3) of [18]:

> 'Hence if the largest component of $\boldsymbol{u}_k$ is the rth, then it is the rth equation which should be omitted when computing $\boldsymbol{u}_k$. This result is instructive but not particularly useful, since we will not know *a priori* the position of the largest component of $\boldsymbol{u}_k$.'

Ipsen, in a very readable survey attributes the idea of omitting one equation of the system to Wilkinson, see Section 7 of [9], but we suspect that this method was routinely taught in mathematics classes *before* Wilkinson was born, see [8], [2], and [14]. He was born in 1919 and [8] was published in 1921.

Wilkinson abandoned the hunt for a good value of r and used $\boldsymbol{b} = PL\boldsymbol{e}$ where $T-\sigma I = PLU$ denotes triangular factorization with partial pivoting and $\boldsymbol{e} = \sum_{i=1}^{n} \pm \boldsymbol{e}_i$, see [17]. However even this choice fails if some eigenvalues are equal to working accuracy and he resorted to 'tweaking' the computed eigenvalues in such cases.

In private communication to one of us Wilkinson declared that he would prefer $\boldsymbol{b} = \boldsymbol{e}_r$ to $\boldsymbol{b} = PL\boldsymbol{e}$ if only he knew a quick, reliable way to choose r so that the rth entry of the wanted vector is above average, not necessarily the greatest.

The current LAPACK codes, see [4], do not use Wilkinson's choice; instead $\boldsymbol{b}$ is chosen 'at random' from an appropriate distribution but this makes it difficult to obtain orthogonal eigenvectors for close eigenvalues. The case for this approach is made in [10].

In this paper we present a new way to choose r that depends strongly on T and σ. However it is not free; the cost is essentially n extra divisions. It turns out that our method produces further information, beyond the right value of r, that helps us avoid the computation of completely negligible entries in the wanted eigenvector. In this way the overhead for finding the right value of r pays for itself as n becomes large. See Figure 4 after reading Section 3.

Our method uses two complete triangular factorizations, one starts from the top and the other from the bottom. This idea, of itself, is not new and forms the basis of 'twisted LDU'. What has not been noticed before is that by combining both sets of 'pivots' one finds the redundancy measure of each row. Then one is in a good position to choose r. Twisted factorization, in contrast, stops the eliminations when they meet at some predetermined interior row. By completing the up and down factorizations at a total cost of $2n$ divisions we have full information on all possible twisted factorizations each of which costs n divisions. A few historical remarks on twisted LDU are given at the end of Section 3.

Section 2 discusses the 'obvious' solution to the problem and shows its shortcomings. The new method is implicit in Theorem 1 which is established in Section 3 along with Theorem 2 which presents accurate ways to compute the determinant. Section 4 shows how the quantities introduced in Theorem 1 reveal the 'envelope' of an eigenvector when the tridiagonal is normal. Section 5 extends the results to cover breakdown in triangular factorization and zero entries in eigenvectors. Section 6 extends the results of Section 3 to block tridiagonal matrices. Application of these ideas and error analysis will be given elsewhere.

The reader is expected to know the LDU theorem concerning existence and uniqueness of triangular factorization and the expressions for the pivots, as the diagonal entries of D are often called. In this representation both L and U have 1's on the diagonal. In practice, when division is slow, people often use $(LD)D^{-1}(DU)$ instead of LDU but the distinction is not important in this paper.

The main notational issue is the representation of submatrices. In MATLAB notation the submatrix of M in rows i through j and columns k through l is given by $M(i:j,k:l)$. This is clear but sometimes too obtrusive. We use $M^{i:j}$ to denote the principal submatrix $M(i:j,i:j)$. For column vectors we prefer simple lower case Latin letters $\boldsymbol{x}$, $\boldsymbol{y}, \ldots$ in bold face type, with entries $x(1), x(2), \ldots, x(n)$. For subvectors we use either $\boldsymbol{x}(i:j)$ or $\boldsymbol{x}^{i:j}$. Finally we try to use lower case Greek letters $\alpha, \beta, \ldots$ for scalars although matrix entries will be written as $M(i,j)$ or M_{ij}.

One notational innovation is to use $+$ to indicate a process taking rows in increasing order and $-$ to indicate the process going in decreasing order, e. g. LDU is written as $L_+D_+U_+$ while UDL is written as $U_-D_-L_-$.

As usual $\|\boldsymbol{v}\| = \|\boldsymbol{v}\|_2 = \sqrt{\boldsymbol{v}^*\boldsymbol{v}}$, while $\|\boldsymbol{v}\|_\infty = \max_i |v(i)|$. The dimension, or order, of $\boldsymbol{e}_1$ or of any column of the identity matrix $I = (\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n)$ is given by the context.

Theorem 1 was presented at the SIAM conference on Parallel Processing in San Francisco, California in February 1995 by two of us (KVF and BNP). Previous work that used a different method to compute the $\boldsymbol{z}^{(k)}$ of Theorem 1, a technique more prone to overflow, was presented (by KVF and BNP) at the Householder XII conference at Lake Arrowhead, California in June 1993, and the SIAM Applied Linear Algebra meeting in Snowbird, Utah in June 1994.

2 A Classical Analysis

In case a pure mathematician should, by chance, read this material it seems wise to begin by explaining that the problem discussed here is not as trivial as it may appear at first. It is the computer's limited precision that causes the difficulties.

Anyone who has mastered an introductory course in matrix theory and who has absorbed the significance of the tridiagonal form J (with nonzero values adjacent to the diagonal) might reason as follows.

Lemma 1 *An eigenvector of an unreduced tridiagonal matrix J cannot have a 0 in the first or last component.*

Proof. Consider the equation for an eigenvector $\boldsymbol{x}$ ($\neq 0$) associated with an eigenvalue λ,

$$(J - \lambda I)\boldsymbol{x} = \mathbf{0}. \tag{2}$$

Suppose that $x(1) = 0$. Then the first equation in (2) dictates that $x(2) = (\lambda - J_{11})x(1)/J_{12} = 0$ as well, since $J_{12} \neq 0$. Now the second equation dictates that $x(3)$ is a linear combination of $x(1)$ and $x(2)$ and also vanishes. Proceeding with the remaining equations, in order, it appears that every entry of $\boldsymbol{x}$ must vanish in contradiction to the assumption that $\boldsymbol{x}$ is an eigenvector. So the assumption that $x(1) = 0$ is not tenable. By similar reasoning but taking the equations in reverse order it is untenable that $x(n) = 0$. □

The preceding argument also shows one way to compute an eigenvector of J. It is valid to set $x(1) = 1$ and to use the first equation of (2) to determine $x(2)$, and the second to determine $x(3)$, using $x(1)$ and $x(2)$. Proceeding as before the rth equation may be used to determine $x(r+1)$ and thus $\boldsymbol{x}$ may be obtained without actually making use of the nth equation which, says the mathematician, will be satisfied automatically since the system (2) is singular.

It would be equally valid to begin with $\tilde{x}(n) = 1$ and to take the equations in reverse order to compute $\tilde{x}(n-1), \ldots, \tilde{x}(2), \tilde{x}(1)$ in turn without using the first equation in (2). When normalized in the same way $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$ will yield the same eigenvector. Note that the problem has been solved without the bother of computing a triangular factorization.

The proof of Lemma 1 actually shows a little more than was claimed. For an upper Hessenberg matrix ((i,j) entry vanishes if $i > j+1$) that is unreduced (entries $(i+1, i)$ do not vanish) $x(n)$ cannot vanish and for an unreduced lower Hessenberg matrix $x(1)$ cannot vanish.

The method described above was proposed by W. Givens in 1954, see [5]. It often gives good results when realized on a computer but, at other times, delivers vectors pointing in completely wrong directions.

The preceding analysis is valid in exact arithmetic but is inapplicable to computer work for the following reasons. First, it is rare that an eigenvalue of a tridiagonal (or any other) matrix is representable in limited precision. Consequently the systems such as (2) that are to be solved are not singular and, in (2), the unused equation will not be satisfied automatically even if the solutions of the other equations, in turn, were obtained exactly. The second weakness is that, in a computer, the sequence $1, x(2), \ldots,$ can overflow. This is a possibility that pure mathematicians do not have to worry about.

It turns out that, for isolated eigenvalues, Givens' method gives an excellent approximate eigenvector whenever the first or last entry of the wanted eigenvector is above average in magnitude. Conversely it gives disastrous results when those extreme entries are tiny. Wilkinson gives a striking example in Section 52, Chap. 5 of [18].

The purpose of this section was to show that the 'obvious' method for computing eigenvectors is not adequate for finite precision arithmetic.

3 Diagonal of the Inverse

In basic courses in matrix theory one is taught to solve a system of equations by computing a row echelon form. If the system is singular at least one row of the echelon form vanishes and the corresponding row of the original system is redundant. The homogeneous system is solved by assigning any values to the 'free' variables and backsolving for the rest of them. In general a discarded row is not unique; it need only be a linear combination of the remaining ones.

In practice our system is nearly, but not quite, singular and a natural modification of the standard procedure is to seek a row that is most nearly redundant and then ignore it while determining a solution $\boldsymbol{x}$ to the *remaining* homogeneous system. This solution $\boldsymbol{x}$ will not satisfy the omitted (rth) equation. In other words, faced with the fact that $M\boldsymbol{x} = \mathbf{0}$ admits only the trivial solution one finds a suitable r and solves, instead, $M\boldsymbol{x} = \boldsymbol{e}_r\delta_r$ where δ_r is the 'defect' or residual of the rth equation.

This is what is meant by 'omitting the rth equation'.

In general it is difficult to find r and to solve the reduced homogeneous system. Fortunately when M is tridiagonal the omission of row r splits the system into two separate parts. For a modest cost the residual δ_j, *for every* *choice of j*, can be computed and that gives an excellent basis for choosing the right r.

Theorem 1 (Double Factorization) *Let J be a tridiagonal $n \times n$ complex matrix that permits triangular factorization in both increasing and decreasing order of rows:*

$$L_+ D_+ U_+ = J = U_- D_- L_-. \tag{3}$$

For each k, $1 \leq k \leq n$, define γ_k and $\boldsymbol{z}^{(k)}$ by

$$J\boldsymbol{z}^{(k)} = \boldsymbol{e}_k \gamma_k, \quad z^{(k)}(k) = 1. \tag{4}$$

Then

$$\gamma_k = D_+(k) + D_-(k) - J_{kk}. \tag{5}$$

Proof. In what follows MATLAB notation will be used for submatrices that are not square and a more condensed representation otherwise. In addition, if terms that involve out of range indices are dropped then the analysis that follows covers the extreme cases $k = 1$ and $k = n$ as well. For brevity write $\boldsymbol{z}$ for $\boldsymbol{z}^{(k)}$.

Omit the kth equation from (4) and what remains is two homogeneous systems. Next use the appropriate triangular factorization (3) to write these systems as

$$\begin{aligned} L_+^{1:k-1} D_+^{1:k-1} U_+(1:k-1, 1:k)\boldsymbol{z}(1:k) &= \mathbf{0}, & (6)\\ U_-^{k+1:n} D_-^{k+1:n} L_-(k+1:n, k:n)\boldsymbol{z}(k:n) &= \mathbf{0}. & (7)\end{aligned}$$

By the assumption that the LDU and UDL factorizations exist the matrices $L_+^{1:k-1}$, $D_+^{1:k-1}$, $U_-^{k+1:n}$, $D_-^{k+1:n}$ must be invertible. Premultiply (6) and (7) by the appropriate inverses to find

$$\begin{aligned} U_+(1:k-1, 1:k)\boldsymbol{z}(1:k) &= \mathbf{0}, & (8)\\ L_-(k+1:n, k:n)\boldsymbol{z}(k:n) &= \mathbf{0}. & (9)\end{aligned}$$

The last equation in (8) shows that

$$1 \cdot z(k-1) + U_{k-1,k}^+ z(k) = 0. \tag{10}$$

The first equation in (9) shows that

$$L_{k+1,k}^- z(k) + 1 \cdot z(k+1) = 0. \tag{11}$$

Recall that $z(k) = 1$ and substitute, from (10) and (11), the values for $z(k-1)$ and $z(k+1)$ into the kth equation of (4) to find, for $k = 2, \ldots, n-1$,

$$\begin{aligned} \gamma_k &= -J_{k,k-1}U^{+}_{k-1,k} + J_{k,k} - J_{k,k+1}L^{-}_{k+1,k} & (12) \\ &= (J_{kk} - J_{k,k-1}U^{+}_{k-1,k}) - J_{k,k} + (J_{kk} - J_{k,k+1}L^{-}_{k+1,k}) \\ &= D_{+}(k) - J_{kk} + D_{-}(k), & (13) \end{aligned}$$

as claimed. For $k = 1$ note that $D_{+}(1) = J_{1,1}$ and $\gamma_1 = D_{-}(1)$. For $k = n$ note that $D_{-}(n) = J_{nn}$ and $\gamma_n = D_{+}(n)$. Thus (5) holds for $k = 1$ an $k = n$ as well as for $k = 2, \ldots, n-1$. □

Corollary 1 *Let J satisfy the Hypotheses of Theorem 1. Either J is singular and then $D_{+}(n) = D_{-}(1) = \gamma_n = \gamma_1 = 0$ and both $\mathbf{z}^{(1)}$ and $\mathbf{z}^{(n)}$ are in J's null space, or*

$$diag(J^{-1})^{-1} + diag(J) = D_{+} + D_{-}\ . \qquad (14)$$

Proof. By the assumption of (3) the only D values that can vanish are $D_{+}(n)$ and $D_{-}(1)$. Also $D_{+}(1) = J_{11}$ and $D_{-}(n) = J_{nn}$ so that, when J is singular

$$\gamma_n = 0 - J_{nn} + D_{-}(n) = 0, \quad \gamma_1 = D_{+}(1) - J_{11} + 0 = 0,$$

and so $J\mathbf{z}^{(1)} = J\mathbf{z}^{(n)} = \mathbf{0}$.

If J is invertible then γ_k must be nonzero for all $k = 1, \ldots, n$ (to avoid giving a nontrivial solution to $J\mathbf{x} = \mathbf{0}$) and multiplication of (4) by $\gamma_k^{-1}J^{-1}$ yields

$$\gamma_k^{-1} = \gamma_k^{-1}z^{(k)}(k) = \mathbf{e}_k^{*}\gamma_k^{-1}\mathbf{z}^{(k)} = \mathbf{e}_k^{*}J^{-1}\mathbf{e}_k. \qquad (15)$$

Thus

$$\gamma_k = \frac{1}{(J^{-1})_{kk}}, \quad k = 1, \ldots, n.$$

□

Equation (14) is a striking property of invertible tridiagonals and gave us the title for this section.

In applications it is useful to have several different expressions for γ_k in addition to (13).

Corollary 2 *With the notation of the Theorem 1, for* $1 < k < n$*,*

$$\gamma_k = \begin{cases} D_+(k) - J_{k,k} + D_-(k), \\ -L^+_{k,k-1} J_{k-1,k} + J_{k,k} - J_{k+1,k} U^-_{k,k+1}, \\ -J_{k,k-1} U^+_{k-1,k} + J_{k,k} - J_{k,k+1} L^-_{k+1,k}, \\ D_+(k) - U^-_{k,k+1} J_{k+1,k}, \\ -L^+_{k,k-1} J_{k-1,k} + D_-(k). \end{cases}$$

For $k = 1$ *and* $k = n$ *omit terms with invalid indices.*

Proof. The first and third expression are just (5) and (12). The others come from rewriting (12) as

$$\gamma_k = -J_{k,k-1} J_{k-1,k} / D_+(k-1) + J_{k,k} - J_{k,k+1} J_{k+1,k} / D_+(k+1) \qquad (16)$$

and using $J_{k,k+1} = U^-_{k,k+1} D_-(k+1) = D_+(k) U^+_{k,k+1}$ etc. and $D_-(k) = J_{kk} - J_{k,k+1} J_{k+1,k} / D_-(k+1)$, etc. □

When J is nearly singular one is most interested in the values of k that yield minimal $|\gamma_k|$ values.

The middle formula in Corollary 2 is of most interest to us because of the following result which shows that no divisions are needed to find $\boldsymbol{z}^{(k)}$ once k is known.

Corollary 3 *With the notation of the Theorem 1 and a given value of k, so that* $\boldsymbol{z} = \boldsymbol{z}^{(k)}$*,* $J\boldsymbol{z} = \boldsymbol{e}_k \gamma_k$*, then*

$$\begin{aligned} z(j) &= -U^+_{j,j+1} z(j+1), \quad j = k-1, \ldots, 1, \\ z(i) &= -L^-_{i,i-1} z(i-1), \quad i = k+1, , \ldots, n. \end{aligned}$$

Proof. These equations are (8) and (9) in expanded form. □

Another reward for computing both factorizations is a wide choice of expressions for *det J*.

Theorem 2 *Assume the hypothesis of Theorem 1. Then for $k = 1, \ldots, n$,*

$$det\, J = \begin{cases} D_+(1)\cdots D_+(k-1)\gamma_k D_-(k+1)\cdots D_-(n) \\ \\ D_+(1)\cdots D_+(k-2)\, det \begin{bmatrix} D_+(k-1) & J_{k-1,k} \\ J_{k,k-1} & D_-(k) \end{bmatrix} D_-(k+1)\cdots D_-(n) \end{cases}$$

and

$$\frac{\gamma_k}{\gamma_{k+1}} = \frac{D_+(k)}{D_-(k+1)}.$$

Proof. Apply Cramer's rule for the kth entry of $\boldsymbol{z}^{(k)}$ where $J\boldsymbol{z}^{(k)} = \boldsymbol{e}_k\gamma_k$. The numerator is a determinant whose kth column is $\boldsymbol{e}_k\gamma_k$. Expand it by column k to find

$$1 = z^{(k)}(k) = \gamma_k\, det\, J^{1:k-1}\, det\, J^{k+1:n}/det\, J.$$

Since $J = L_+D_+U_+ = U_-D_-L_-$ it follows that

$$det\, J^{1:k-1} = D_+(1)\cdots D_+(k-1), \quad det\, J^{k+1:n} = D_-(k+1)\cdots D_-(n).$$

The second expression comes from the twisted factorization of J:

$$J = \begin{bmatrix} L_+^{1:k-1} & O \\ O & U_-^{k:n} \end{bmatrix} \begin{bmatrix} D_+^{1:k-2} & & \\ & \Box & \\ & & D_+^{k+1:n} \end{bmatrix} \begin{bmatrix} U_+^{1:k-1} & O \\ O & L_-^{k:n} \end{bmatrix}$$

where

$$\Box = \begin{bmatrix} D_+(k-1) & J_{k-1,k} \\ J_{k,k-1} & D_-(k) \end{bmatrix}.$$

From the first expression for *det* J it follows that

$$\gamma_k D_-(k+1) = \gamma_{k+1} D_+(k),$$

which gives the ratio of consecutive γ's. □

When there is severe cancellation in computing γ_k from any of the formulae in Corollary 2 then it may be possible to take extra care in the evaluation of $det\,\square$ and win back a few bits of precision. If warranted the idea may be taken further to use

$$det\ tridiag \begin{bmatrix} D_+(k-1) & J_{k-1,k} & 0 \\ J_{k,k-1} & J_{k,k} & J_{k,k+1} \\ 0 & J_{k+1,k} & D_-(k+1) \end{bmatrix}$$

for the sensitive part of the computation. These details are of great practical importance when J is close to singular as occurs in iterative methods for finding eigenvalues.

Remark 1 An attentive reader may be puzzled that Corollary 3 cannot generate an isolated 0 entry in $\boldsymbol{z}^{(k)}$. If $z(j)$ vanishes (because $J(j,j+1)=0$) then all entries $z(l)$, $l<j<k$, must vanish too. This is appropriate since the matrix is reduced when $J_{j,j+1}=0$. Yet there exist eigenvectors of unreduced tridiagonals with isolated entries that vanish; such entries are the nodes of the eigenvector. The explanation is that the hypothesis (13) does indeed rule out isolated zero entries. Section 5 extends the results of Theorem 1 to cover these important cases. There we see that hypothesis (3) in theorem 1 is not essential.

Remark 2 Corollaries 2 and 3 show that we need only retain the nontrivial entries of L_- and U_+ in the factorization process in order to obtain all the γ-values, and for any given r, to solve for $\boldsymbol{z}^{(r)}$ with no more divisions.

Remark 3 For large n there will be many products in the calculation of $z(1)$ or $z(n)$ for a given r. In general one is concerned about possible overflow but here r is selected so that $z(r)$ should be a maximal, or nearly maximal, entry of $\boldsymbol{z}$ and so no overflow can occur. Underflow, if it occurs, is harmless here and should be flushed to 0.

Remark 4 From (4) we see that the FP vector $\boldsymbol{z}^{(r)}$ is annihilated exactly by $J-\boldsymbol{e}_r\gamma_r\boldsymbol{e}_r^*$, a rank one perturbation.

Let us summarize this section in the form of a high level algorithm.

Null Vector Approximation.

Input: J, nearly singular, $n \times n$, tridiagonal

1. Factor J down and up to compute $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)$ using a convenient formula from Corollary 2.
2. Find $\min_i |\gamma_i|$, perturbing zero values appropriately. Compute the index set $\mathcal{L} := \{j : |\gamma_j| < \mu \min_i |\gamma_i| \}$. Choose μ so that $1 < \mu < 2$.
3. Select a suitable r from $\mathcal{L}$ and compute $\boldsymbol{z}^{(r)}$ using Corollary 3. It is also possible to take an appropriate linear combination of two or more $\boldsymbol{z}^{(j)}$, $j \in \mathcal{L}$.

Output: r and $\boldsymbol{z}^{(r)}$.

We do not claim that $\boldsymbol{z}^{(r)}$, which we call the FP vector, is always an adequate approximation to a null vector of J. However it is always useful. We do not discuss the calculation of orthogonal vectors for clustered eigenvalues here, see [13]. It seems to be wise, when eigenvalues are real, to forbid the same r to adjacent eigenvalues.

There are several possibilities for replacing zero values of γ_k in Step 2. Theorem 2 gives one way and Remark 4 suggests that we set $\gamma_k = macheps \cdot J_{kk}$.

Estimates of the accuracy of $\boldsymbol{z}^{(r)}$ as an approximate null vector may be given in different contexts: the nonsymmetric case, the real symmetric case, the general linear eigenvalue problem, and the bidiagonal singular vector case. We do not wish to submerge the ideas in this paper with such results but see the corollary to Theorem 3 in Section 4.

As mentioned in the introduction twisted LDU starts elimination from the top and the bottom and stops at some selected interior row k. Henrici (1963) used twisted LDU implicitly in deriving optimal Gersgorin bounds, in [7], for unreduced real tridiagonals with some complex eigenvalues. D. Kershaw (1970), in [11], obtained nice bounds on $(J^{-1})_{rr}/J_{rr}$ using twisted LDU. Babuska (1972) wanted a specific entry in $J^{-1}\boldsymbol{b}$ for any $\boldsymbol{b}$ and his formula (5.29) on page 62 of [1] is one instance of our formula (5). Fischer et. al. (1974), in [6], discuss a twisted Toeplitz factorization of Buneman and attribute the adjective 'twisted' to Strang [15]. Dongarra et. al. (1979), in the LINPACK codes use twisted LDU meeting in the middle for improved efficiency and the practice has been taken up in parallel computation, see

[16]. They call it the BABE algorithm (Begin, or Burn, At Both Ends), see [3].

4 The Eigenvector Connection

Suppose that $\tilde{J}$ is a normal matrix as well as tridiagonal;

$$\tilde{J} = V\Lambda V^*, \quad V^{-1} = V^*,$$

and

$$\lambda = diag(\lambda_1, \ldots, \lambda_n), \quad V = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n).$$

Theorem 3 *Let $J = \tilde{J} - \sigma I$ satisfy the hypotheses of Theorem 1. Then $\gamma_k = \gamma_k(\sigma)$, for each k, and as $\sigma \longrightarrow \lambda_j$, where λ_j is an isolated eigenvalue of $\tilde{J}$, then*

$$\frac{\gamma_k^{-1}}{\sum_{m=1}^n \gamma_m^{-1}} \longrightarrow |v_j(k)|^2,$$

for all k, $1 \le k \le n$, such that $v_j(k) \neq 0$.

Proof. From (15), for each k

$$\begin{aligned}
\gamma_k^{-1} &= \boldsymbol{e}_k^*(\tilde{J} - \sigma I)^{-1}\boldsymbol{e}_k \\
&= \boldsymbol{e}_k^* V(\Lambda - \sigma I)^{-1} V^* \boldsymbol{e}_k \\
&= \sum_{i=1}^n \frac{|V_{ki}|^2}{(\lambda_i - \sigma)} \\
&= (\lambda_j - \sigma)^{-1}\left\{ |v_j(k)|^2 + \sum_{i \neq j} |v_i(k)|^2 \left(\frac{\lambda_j - \sigma}{\lambda_i - \sigma}\right)\right\}.
\end{aligned}$$

Sum for $k = 1, \ldots, n$ and use the orthogonality of V to find

$$\begin{aligned}
\sum_{k=1}^n \gamma_k^{-1} &= (\lambda_j - \sigma)^{-1}\left\{1 + \sum_{i \neq j}\left(\frac{\lambda_j - \sigma}{\lambda_i - \sigma}\right) \sum_{k=1}^n |v_i(k)|^2\right\} \\
&= (\lambda_j - \sigma)^{-1}\left\{1 + \sum_{i \neq j}\left(\frac{\lambda_j - \sigma}{\lambda_i - \sigma}\right)\right\}.
\end{aligned}$$

Hence, as $\sigma \longrightarrow \lambda_j$, an isolated eigenvalue,

$$\begin{aligned}\frac{\gamma_k^{-1}}{\sum_{m=1}^n \gamma_m^{-1}} &= \left\{|v_j(k)|^2 + \sum_{i\neq j} |v_i(k)|^2 \left(\frac{\lambda_j-\sigma}{\lambda_i-\sigma}\right)\right\}\left\{1+\sum_{i\neq j}\left(\frac{\lambda_j-\sigma}{\lambda_i-\sigma}\right)\right\}^{-1} \\ &\longrightarrow |v_j(k)|^2,\end{aligned}$$

provided that $|v_j(k)|^2 > 0$. □

Since

$$\sum_{i\neq j} |v_i(k)|^2 = 1 - |v_j(k)|^2$$

convergence is more rapid the larger is $|v_j(k)|$. For a given σ, much closer to λ_j than to any other λ_i, the vector with components $(\gamma_k^{-1}/\sum_{m=1}^n \gamma_m^{-1})^{1/2}$, $k = 1, \ldots, n$, provides a useful envelope of the eigenvector $\boldsymbol{v}_j$. It is pleasing that there is no requirement that the γ_k be real; the appropriate quotients will be positive.

The FP vector $\boldsymbol{z}$ defined by (4) is an alternative approximation to the eigenvector $\boldsymbol{v}_j$ of $\tilde{J}$. Its quality is indicated in the next result.

Corollary 4 *With the notation of Theorem 3 let* $J\boldsymbol{z} = (\tilde{J} - \sigma I)\boldsymbol{z} = \boldsymbol{e}_k\gamma_k$. *Then*

(a) Rayleigh quotient$(\boldsymbol{z}) = \sigma + \gamma_k/\|\boldsymbol{z}\|^2$,

(b) $|\lambda_j - \sigma - \gamma_k/\|\boldsymbol{z}\|^2| \leq |\gamma_k|/\|\boldsymbol{z}\| \cdot \min\{1, |\gamma_k|/(\|\boldsymbol{z}\|gap)\}$,
where $gap = \min\{|\lambda_{j+1} - \sigma|, |\sigma - \lambda_{j-1}|\}$,

(c) $\sin \angle(\boldsymbol{v}_j, \boldsymbol{z}) \leq |\gamma_k|/(\|\boldsymbol{z}\|gap)$.

Proof. $\boldsymbol{z}^*\tilde{J}\boldsymbol{z}/\boldsymbol{z}^*\boldsymbol{z} = \sigma + \boldsymbol{z}^*\boldsymbol{e}_k\gamma_k/\|\boldsymbol{z}\|^2$, gives (a). Also $\|J\boldsymbol{z}\|/\|\boldsymbol{z}\| = |\gamma_k|/\|\boldsymbol{z}\|$ and standard gap theorems, see Chapter 11 in [12], give (b) and (c). □

Unfortunately, in finite precision arithmetic, we cannot let $\sigma \longrightarrow \lambda_j$ and so, in practice, a γ_j may vanish in the same way that a divided difference may vanish even when the desired derivative does not. However we do not need more than a few bits of accuracy in γ_j, it is its exponent that counts.

Figure 1 shows the γ-vector for different values of σ and the true profile of a simple eigenvector, all on a log scale.

Theorem 4 *An unreduced tridiagonal matrix is normal if, and only if, it is a translate of a Hermitian or skew-Hermitian matrix, possibly multiplied by a rotation factor $e^{i\theta}$, $i^2 = -1$.*

The proof is left to the interested reader.

5 Zero Pivots

Triangular factorization is said to fail, or not exist, if a zero 'pivot', $D_+(j)$ or $D_-(j)$ is encountered prematurely. The last pivot is allowed to vanish because it does not occur as a denominator in the computation.

One of the attractions of an unreduced tridiagonal matrix is that the damage done by a zero pivot is localized. Indeed, if ∞ is added to the number system then triangular factorization cannot break down and the algorithm always maps J into unique triplets L, D, U. There is no need to spoil the inner loop with tests. It is no longer true that $LDU = J$ but equality does hold for all entries except for those at or adjacent to any infinite pivot.

It is possible to work with signed ∞ (affine geometry) or unsigned ∞ (the complex plane) and it will be easiest for our purposes to use the unsigned ∞. Thus $+1/0 = -1/0 = \infty$.

If we allowed off diagonal entries to vanish, in which case J is said to be *reduced*, then we might encounter

$$L(k+1,k) = J(k+1,k)/D(k) = 0/0$$

and that would be a genuine breakdown.

Let us examine what happens when $D(k-1) = 0$. In turn

$$\begin{aligned}
L(k,k-1) &= J(k,k-1)/D(k-1) = \infty, \\
\bullet \quad U(k-1,k) &= J(k-1,k)/D(k-1) = \infty, \\
D(k) &= J(k,k) - L(k,k-1)J(k-1,k) = \infty, \\
L(k+1,k) &= J(k+1,k)/D(k) = 0, \\
\bullet \quad U(k,k+1) &= J(k,k+1)/D(k) = 0, \\
D(k+1) &= J(k+1,k+1) - L(k+1,k)J(k,k+1) \\
&= J(k+1,k+1).
\end{aligned}$$

Unless $J(k+1,k+1)=0$ the factorization proceeds normally until the next zero pivot is encountered. We have placed an $\bullet$ against entries that are not computed when a simple $L\tilde{U}$ factorization is used. Here $\tilde{U} = DU$ in the finite case.

When the product LDU is formed in the case given above then various strange expressions such as $0 \cdot \infty$ and $\infty + \infty$ arise and we designate them by NaN (Not a Number). We discover that $LDU = J$ except in row and column k. Note that $D(k) = \infty$.

It is important to our later results to show that when J is singular then

$$D_+(k) = \infty \quad \text{if, and only if,} \quad D_-(k) = \infty,$$

where the notation follows Section 3.

It turns out that infinite pivots correspond to zero entries in eigenvectors and so have a legitimate role in the theory.

Theorem 5 *Let J be $n \times n$, tridiagonal, unreduced, and singular. For each k, $1 < k < n$, $J^{1:k-1}$ is singular if, and only if, $J^{k+1:n}$ is singular. They are singular if, and only if, $z(k) = 0$ whenever $J\boldsymbol{z} = \boldsymbol{0}$.*

Proof. Write

$$\boldsymbol{z} = \begin{pmatrix} \boldsymbol{z}_+ \\ z(k) \\ \boldsymbol{z}_- \end{pmatrix}$$

and partition $J\boldsymbol{z} = \boldsymbol{0}$ conformably. Thus

$$J^{1:k-1}\boldsymbol{z}_+ + J_{k-1,k}z(k)\boldsymbol{e}_{k-1} = \boldsymbol{0}, \tag{17}$$

$$\boldsymbol{e}_1 J_{k+1,k}z(k) + J^{k+1:n}\boldsymbol{z}_- = \boldsymbol{0}, \tag{18}$$

and $\boldsymbol{z}_+(1) \neq 0$, $\boldsymbol{z}_+(n) \neq 0$ by Lemma 1 in Section 2.

If $z(k) = 0$ then (17) shows that $\boldsymbol{z}_+ (\neq 0)$ is in $J^{1:k-1}$'s null space and (18) shows that $\boldsymbol{z}_- (\neq 0)$ is in $J^{k+1:n}$'s null space. So both matrices are singular.

Now consider the converse, $z(k) \neq 0$. Since J is unreduced $rank(J) = n - 1$ and its null space is one dimensional. So the system

$$J\boldsymbol{z} = \boldsymbol{0}, \quad z(k) = 1,$$

has a unique solution. Thus both (17) and (18) are inhomogeneous equations with unique solutions. Thus $J^{1:k-1}$ and $J^{k+1:n}$ are invertible. □

Corollary 5 *Let J be $n \times n$, tridiagonal, unreduced, and singular. Let the triangular factorization algorithm applied to J in both increasing and decreasing order of rows yield unique matrices L_+, D_+, U_+, U_-, D_-, and L_-. Then, for $j = 1, 2, \ldots, n$,*

$$D_+(j) = \infty \quad \textit{iff} \quad D_-(j) = \infty.$$

Proof.

$$\begin{aligned} D_+(j) = \infty &\iff D_+(j-1) = 0 \\ &\iff J^{1:j-1} \ \text{ singular} \\ &\iff J^{j+1:n} \ \text{ singular (by Theorem 5)} \\ &\iff D_-(j+1) = 0 \\ &\iff D_-(j) = \infty. \end{aligned}$$

□

In Theorem 1 of Section 3 the value of γ_k was fixed by the condition $z(k) = 1$ imposed on the solution of $J\boldsymbol{z} = \boldsymbol{e}_k \gamma_k$. When J is singular there is a nonzero solution to $J\boldsymbol{z} = \boldsymbol{0}$ and the attempted normalization $z(k) = 1$ is valid, even if not wise, in all cases *except when* $z(k) = 0$.

An appropriate signal that an infeasible normalization has been imposed is that $\gamma_k = \text{NaN}$ (Not a Number) and that is precisely what the formulae in Corollary 2 deliver whenever J is singular and $D_+(k) = D_-(k) = \infty$. In these cases, in exact arithmetic, $D_+(n) = 0$ and $\gamma_n = 0$ as well as $D_-(1) = 0$ and $\gamma_1 = 0$. Thus in the search for a minimum value of $|\gamma_j|$ indices j that have $\gamma_j = \text{NaN}$ will never be selected.

The good news is that by computing all the $\{\gamma_j\}$ it is known *in advance* whether or not the $\boldsymbol{z}^{(k)}$ has a zero entry. In the generic case, with no zeros, the algorithm given in Corollary 1 in Section 3 may be used free of any tests for invalid operations. In the exceptional case the following procedure may be used.

Algorithm (vectors with zeros)

$$z(j) = \left\{ \begin{array}{ll} -U^+_{j,j+1} z(j+1), & z(j+1) \neq 0, \\ -(J_{j+1,j+2}\, z(j+2)/J_{j+1,j}), & \text{otherwise} \end{array} \right\}, \; j = k-1, \ldots, 1$$

$$z(i) = \left\{ \begin{array}{ll} -L^{-}_{i,i-1} z(i-1), & z(i-1) \neq 0, \\ -(J_{i-1,i-2}\, z(i-2)/J_{i-1,i}), & \text{otherwise} \end{array} \right\}, \; i = k+1, \ldots, n$$

This algorithm will not touch any infinite values in U_+ or L_-.

When J is not singular Theorem 1 continues to hold, if ∞ is allowed, in the following sense.

Corollary 6 (of Theorem 2) *If J is unreduced, tridiagonal, and ∞ is represented then*

$$\begin{array}{rll} D_+(k-1) = 0 & \text{implies} & (J^{-1})_{kk} = 0 \;\; (\gamma_k = \infty), \\ D_-(k) = 0 & \text{implies} & (J^{-1})_{k-1,k-1} = 0 \;\; (\gamma_{k-1} = \infty). \end{array}$$

Proof. Use the twisted factorization in the proof of Theorem 2 that introduces the 2×2 matrix

$$\Box = \begin{bmatrix} D_+(k-1) & J_{k-1,k} \\ J_{k,k-1} & D_-(k) \end{bmatrix}.$$

Invert J and observe that there is a simple expression for the (k,k) and $(k-1,k-1)$ entries:

$$(J^{-1})_{kk} = (\Box^{-1})_{2,2} = D_+(k-1)/det\,\Box\,.$$

If J is unreduced and $D_+(k-1) = 0$ then $det\,\Box = -J_{k,k-1}J_{k-1,k} \neq 0$. This establishes the first assertion. Similarly

$$(J^{-1})_{k-1,k-1} = D_-(k)/det\,\Box\,.$$

□

Figures 2 and 3 show a striking instance of Theorem 5 for the matrix W^+_{21} and the pair of eigenvalues close to 6. Each horizontal line of the figure corresponds to one value of k; eigenvalues of $W^{1:k-1}$ are marked by $+$ and eigenvalues of $W^{k+1:n}$ are marked by $\circ$. Theorem 5 implies that if an eigenvector has a zero entry in position k then a $\circ$ and a $+$ must coincide on the eigenvalue in line k. Indeed, in Figure 3 (an enlarged picture of Figure 2 near 6), when $k = 11$ this is precisely what happens. For neighboring values of k the Ritz values are not particularly close to eigenvalues and after $k = 11$ the $\circ$ is replaced by a $+$ in the interval $(\lambda_{12}, \lambda_{13})$. If $\boldsymbol{v}$ is a normalized eigenvector with eigenvalue λ then $v(k)^2$ is proportional to the product of the distances of λ from the $+$ and $\circ$ points on line k.

6 Block Tridiagonals

If an arithmetic system lacks the symbol ∞ it is possible to extend Theorem 1 by using blocks in the LDU factorization. If J is unreduced then there always exists a factorization

$$L_+D_+U_+ = J = U_-D_-L_-$$

if the D's are allowed to have 2×2 and 1×1 blocks along diagonal, no larger blocks are needed. However Theorem 1 extends beyond this case to band matrices and, to any block tridiagonal matrix. Thus D_+ and D_- are direct sums of square blocks; L_+ and U_+ are conformable with D_+, L_- and U_- are conformable with D_-.

Theorem 6 *Let J permit block triangular factorization in both increasing and decreasing order of indices*

$$L_+D_+U_+ = J = U_-D_-L_- \, .$$

There is no requirement that the block structures of D_+ and D_- be conformable. However for any corresponding blocks k and l such that $D_+(k)$ and $D_-(l)$ are conformable and m by m, define the $m \times m$ matrix Γ by equations

$$J \begin{pmatrix} Z^+ \\ I_m \\ Z^- \end{pmatrix} = \begin{pmatrix} O \\ \Gamma \\ O \end{pmatrix}, \quad Z = \begin{pmatrix} Z^+ \\ I_m \\ Z^- \end{pmatrix}. \tag{19}$$

If $J_{\cdot,\cdot}$ denotes the $m \times m$ block of J conformable with $D_+(k)$ and $D_-(l)$, then

$$\Gamma = D_+(k) - J_{\cdot,\cdot} + D_-(l).$$

The proof is so similar to the proof of Theorem 1 that we omit it. We have allowed for the fact that D_+ and D_- need not have the same number of blocks.

To use Theorem 6 to approximate an eigenvector suppose that J is nearly singular. Compute all well defined Γ and find one with a minimal singular value. Call it $\overset{o}{\Gamma}$. Let

$$\overset{o}{\Gamma}\, \boldsymbol{v} = \boldsymbol{u}\sigma_{min}, \quad \|\boldsymbol{u}\| = \|\boldsymbol{v}\| = 1,$$

define the minimal singular triple $(\sigma_{min}, \boldsymbol{u}, \boldsymbol{v})$ of $\overset{o}{\Gamma}$. Then, from (19)

$$J(Z\boldsymbol{v}) = \begin{pmatrix} \mathbf{0} \\ \boldsymbol{u} \\ \mathbf{0} \end{pmatrix} \sigma_{min}.$$

If σ_{min} is small enough then $Z\boldsymbol{v}$ is a good initial approximation to an eigenvector of J.

It is not hard to verify that, for an unreduced even order J, if $diag(J) = 0$ then $diag(J^{-1}) = 0$. In this situation a block factorization with 2 × 2 blocks is needed to ensure that $J = L_+ D_+ U_+ = U_- D_- L_-$. It then turns out that

$$D_+ = D_- = block\ diag(J) = block\ diag(J^{-1})^{-1}$$

where $bdiag(M)$ is the block diagonal part of M. Now the set of Γ matrices in Theorem 6 may give no guidance for computing an eigenvector. That is not quite true because we may infer that our eigenvalue approximation, 0, is not closer to one eigenvalue than to any other and that is useful information. In fact the unreduced J's with $diag(J) = 0$ have eigenvalues in ± pairs and any tiny ± pairs may be found efficiently by the method described in [13].

Acknowledgement
We are glad to acknowledge helpful discussions with the LAPACK group at UC, Berkeley, in particular Ming Gu and Velvel Kahan. We thank Gene Golub for the Babuska reference.

References

[1] I. Babuska, *Numerical stability in problems of linear algebra*, SIAM J. Numer. Anal., 9 (1972), pp. 53–77.

[2] S. H. Crandall and W. G. Strang, *'An improvement of the Holzer table based on a suggestion of Rayleigh's'*, J. Appl. Mech., 24 (1957), pp. 228-230. Also Discussion in 25 (1958), pp. 160–161.

[3] J. J. Dongàrra et al., *LINPACK*, SIAM, Philadelphia, 1979.

[4] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen, *LAPACK Users' Guide, Release 2.0*, SIAM, Philadelphia, 1995.

[5] W. Givens, *'Numerical computation of the characteristic values of a real symmetric matrix'*, Oak Ridge Nat. Lab. Report, ORNL-1574 (1954).

[6] G. H. Golub, D. Fischer, O. Hald, C. Leiva and O. Widlund, *'On Fourier Toeplitz methods for separable elliptic problems'*, Math. Comp., 28 (1974), pp. 349–368.

[7] P. Henrici, *'Bounds for eigenvalues of certain tridiagonal matrices'*, SIAM J., 11 (1963), pp. 281–290.

[8] H. Holzer, *'Die Berechnung der Drehschwingungen'*, Springer-Verlag, Berlin, 1921.

[9] I. Ipsen, *'A history of inverse iteration'*, in Helmut Wielandt, Mathematische Werke, B. Huppert and H. Schneider, eds., vol.II: Matrix Theory and Analysis, Walter de Guyter, Berlin, 1995.

[10] E. Jessup and I. Ipsen, *'Improving the accuracy of inverse iteration'*, SIAM J. Sci. Stat. Comput. 13 (1992), pp. 570–572.

[11] D. Kershaw, *Inequalities on elements of the inverse of a certain tridiagonal matrix*, Math. Comp. 24 (1970), pp. 155–158.

[12] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice-Hall, N. J., 1980.

[13] B. N. Parlett, *The Construction of orthogonal eigenvectors for tight clusters by use of submatrices.* In progress.

[14] R. V. Southwell, *Relaxation Methods in Engineering Science*, Clarendon Press, Oxford, 1940.

[15] W. G. Strang, *'Implicit difference methods for initial boundary value problems'*, J. Math. anal. Appl., 16 (1966), pp. 188–198.

[16] H. A. van der Vorst, *Analysis of a parallel solution method for tridiagonal linear systems*, Parallel Computing, 5 (1987) pp. 303–311.

[17] J. H. Wilkinson, *'The calculation of the eigenvectors of codiagonal matrices'*, Computer J., 1 (1958), pp. 90–96.

[18] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

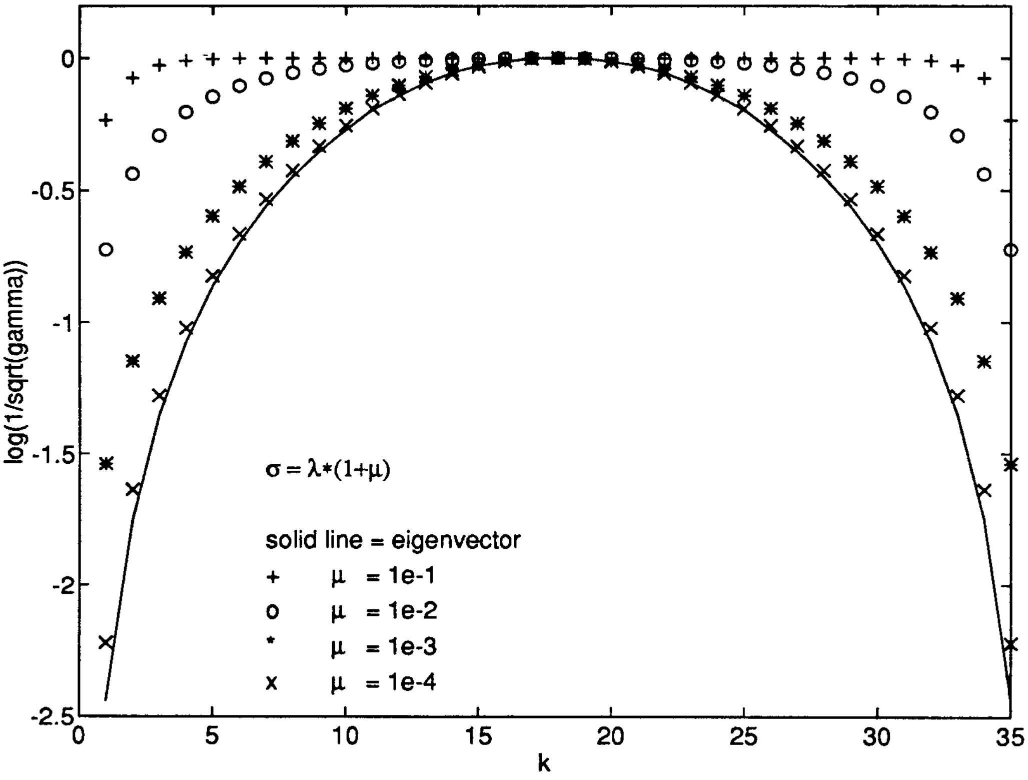


Figure 1: Convergence of $\left(\frac{1}{\gamma}\right)^{1/2}$, $n = 35$

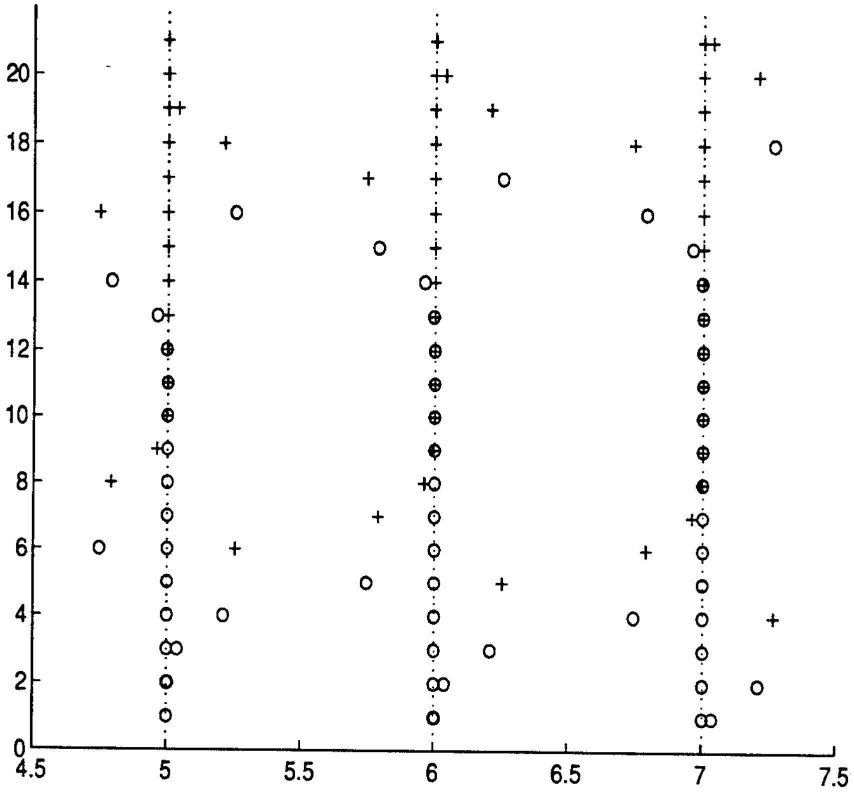


Figure 2: Ritz values for W_{21}^{+} near 6

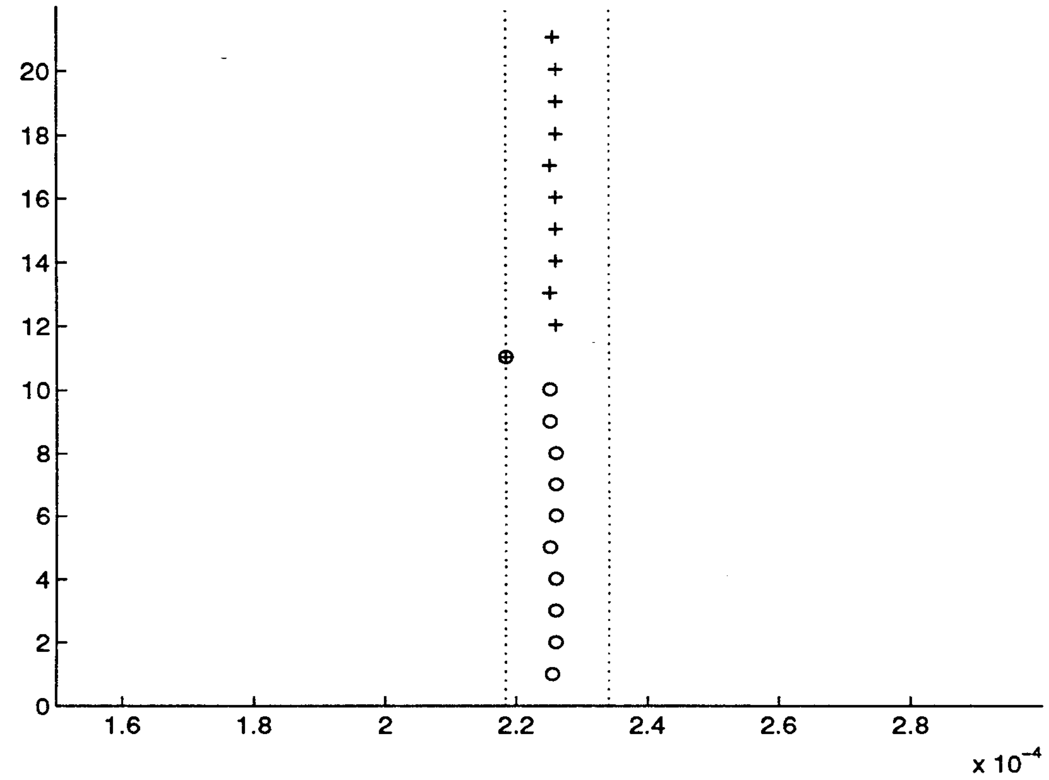


Figure 3: Blow up of Figure 2 near 6 (read the x-axis as 6+x)

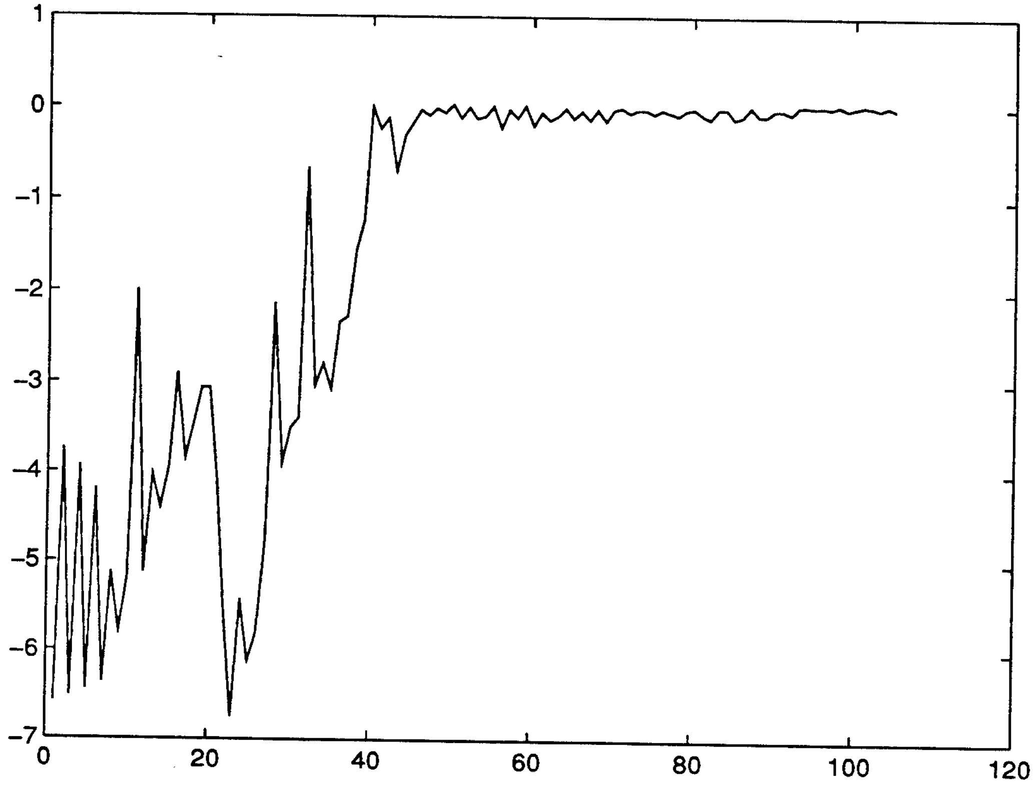


Figure 4: log γ; negligible eigenvector entries from 42 to 105